Collaboratory for Multi-Scale Chemical Science

Status as of March, 2002 / Quarterly Report for Q1 and Q2 of FY 2002

Project Staff

Larry Rahn-SNL*, Christine Yang, John C. Hewson, Carmen Pancerella, Wendy Koegler, David Leahy, Jeff Nichols-PNL*, Brett Didier, Theresa Windus, James D. Myers, Karen Schuchardt, Eric Stephan, Al Wagner-ANL*, Branko Ruscic, Michael Minkoff, Lee Liming, Sandra Bittner, Brian Moran, Gregor von Laszewski, William Pitz-LLNL*, David R. Montoya-LANL*, Thomas C. Allison-NIST*, William H. Green-MIT*, Jr., Michael Frenklach-UCB*

* denotes Institutional Point of Contact

Overview

The pilot Collaboratory for Multi-Scale Chemical Sciences (CMCS) brings together leaders in scientific research and technological development across multiple DOE laboratories, other government laboratories and academic institutions. This group is developing an informatics-based approach to synthesizing multi-scale information, making possible the creation of new scientific methodologies and new knowledge in the chemical sciences. The CMCS will use advanced collaboration and metadata-based data management technologies to develop an MCS (Multi-Scale Chemical Sciences) portal, providing community communications mechanisms and data search and annotation capabilities. The portal will also provide capabilities for defining and browsing cross-scale dependencies between data produced at one scale that is used as input for computations at the next. Notification mechanisms will make both researchers and their applications aware of updated relevant information such as reaction rates. The CMCS and its MCS portal will provide mechanisms to enhance the coordination of research efforts across related sub-disciplines in the chemical sciences, focusing research at one scale on obtaining or refining values critical in the next, reducing work performed using limited or outdated values, and enhancing the ability of the community to meet the national research challenges of the DOE.

Progress

This document summarizes the work done over the first two quarters of the CMCS project. First, a summary listing of project activities is presented, then more a detailed discussion of accomplishments by project area.

Summary of major CMCS Project Activities – Project initiation through March 2002

- Held first CMCS Project Team meeting (7/2001)
- Identified basic infrastructure needs for CMCS project team (7/2001)

- Established initial CMCS web site to provide basic tools for project team (8/2001 – 9/2001)
- Held CMCS WEBShop – a two day web-based poster workshop (9/2001)
- Acquired, upgraded, and installed hardware to support CMCS project (9/2001)
- Developed application use cases to identify project requirements (8/2001 - 10/2001)
    - Active Tables
    - Chemical Kinetic Mechanisms
    - Consensus
    - Feature Tracking
    - Molecular Calculations
    - Portal
    - Reacting Flow
    - Gri-MECH
- Started CMCS Web Seminars series as a mechanism to provide information to team members (10/2001)
- Installed DAV server for CMCS development (10/2001)
- Developed Process/Design Concept documents for key CMCS technology areas (11/2001)
- Held two day CMCS Project Team Workshop (face-to-face) at SNL (11/2001)
- Formed Infrastructure Task Groups to investigate, evaluate and recommend specific technology approaches; assigned task group leads (12/2001)
    - Data/Metadata Management
    - Notification
    - Pedigree
    - Search
    - Personalization/Portal
    - Security
- Published DAV Tutorial (12/2001)
- Published CMCS Management Plan (1/2002)
- Published CMCS Software Engineering Plan (1/2002)
- Developed proposal for Reacting Flow Data Specification (1/2002)
- Participated in SciDAC PI meeting (1/2002)
- Participated in National Collaboratories Workshop (1/2002)
- Drafted CMCS Pedigree Requirements Document (2/2002)
- Formed Prototype Application/Database Communication Task Group (2/2002)
- Drafted document describing CMCS Data Storage Client API (2/2002)
- Participated in Global Grid Forum (2/2002)
- Drafted Use Cases for Security Task Group (3/2002)
- XML Data Schema (3/2002)
- Drafted Notification Requirements Document (3/2002)
- Formed GRI-Mech Plus Development Task Group (3/2002)
- Drafted document describing CMCS Data Storage Interface API Pseudo Code (3/2002)
- Participated in SciDAC Technical Panel on Portals (3/2002)
- Drafted plan for Managing CMCS Jetspeed Development (3/2002)

Project Management Structure

The management structure and processes of the CMCS project has been developed and documented in the CMCS Management Plan. The structure incorporating the individual investigators includes a Project Director, Point of Contact Team (POC), Advisory Board, Chief Technology Officer, Chief Integration Officer, technical working groups, and task groups.

CMCS has three working groups: application, portal, and infrastructure. CMCS working groups are long-term groups for discussion of broad technology/style and discipline-specific issues. These groups also provide a mechanism to focus technical activities and to track progress. Working Groups develop background information, technical approaches to meeting requirements, and assess the success or failure of project technologies and/or processes.

At this point in the project, CMCS Task Groups are oriented towards defining initial functionality and creating prototypes of implementations. The prototypes will be integrated over the summer, and will be presented in demonstrations in the Fall of 2002. The overall goal is to achieve a working, public version of CMCS in this same time frame.

These task groups, which are dynamically formed as needed, are associated with the working groups. For example, task groups that have either been completed, are in progress, or are in conception phase are:

- CMCS POC and Leadership Teams
    - Workshop #1
    - Initial infrastructure
    - Software engineering plan
    - Workshop #2
- Application Task Groups
    - Data model/XML
    - Prototype Database/Application Communication
    - Translation of data from thermochemical to reacting flows
    - Project level tool support
    - Data (mining)/annotation
- Infrastructure Task Groups
    - Data/Metadata Management
    - Search
    - Notification
    - Pedigree
    - Security
- Portal Task Group
    - Portal

Each of these task groups is charged with defining their scope and task timelines, and most are in the progress of completing specific tasks. On an ongoing basis, the Task Groups are coordinated by the CTO, CIO and/or the relevant Working Group lead.

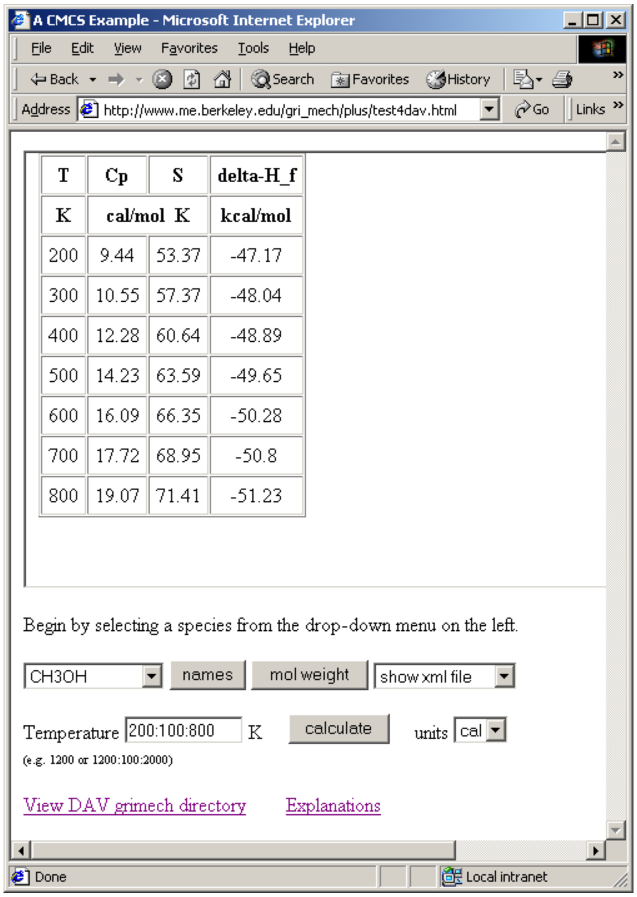
Some task groups have finished entirely, including all of the CMCS POC and Leadership Teams task groups except that for Workshop#2. Other task groups are still in initial stages of heavy overlap as the early design work and implementations are being finished; this is particularly true of the Infrastructure task groups. Specifically, the Data/Metadata Management Task Group, Pedigree Task Group, Search Task Group, Notification Task Group, and Security Task Group are deeply inter-dependent at this early stage, and will develop into more separable tasks as the project evolves.

The following is a discussion of the project progress organized by Working Group.

Application Working Group

The applications researchers have been discussing and documenting several issues related to the CMCS capabilities. Among these issues have been the data model definitions (what data needs to be transferred across scales, what data is useful essentially only at one scale, etc.), data schemas to be used, work flow issues with the portal (these have been documented in the form of use cases and other associated documents), data storage and transfer using the DAV protocol, translation issues between legacy codes and newly developed codes that will use the CMCS data schema, search capabilities within the CMCS and with other public chemical databases, and data/model comparison.

We have been developing the chemical mechanism and reacting flow scale data model, use cases, and initial versions of some of the chemical analysis tools. Prototype implementations of XML have been developed in the context of existing data models (CML, XSIL, Dublin Core) that provide descriptions of species, data sets, and pedigrees. These activities are coupled to applications focusing on the inter-scale sharing of data and feature detection in the analysis of reacting flow data.


A CMCS Example - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://www.me.berkeley.edu/gri_mech/plus/test4dav.html
Go Links
Done
Local intranet

T	Cp	S	delta-H_f
K	cal/mol K		kcal/mol
200	9.44	53.37	-47.17
300	10.55	57.37	-48.04
400	12.28	60.64	-48.89
500	14.23	63.59	-49.65
600	16.09	66.35	-50.28
700	17.72	68.95	-50.8
800	19.07	71.41	-51.23

Begin by selecting a species from the drop-down menu on the left.

CH3OH | names | mol weight | show xml file

Temperature 200:100:800 K | calculate | units cal

(e.g. 1200 or 1200:100:2000)

View DAV grimech directory   Explanations

Figure 1. Web-based viewer displaying methanol thermodynamic data.

A prototyping task group has completed a demonstration (Figure 1) of reading XML data from the CMCS WebDAV server into a thermochemistry viewer. The GRI-Mech thermochemistry data has been represented in an XML scheme that was parsed in JavaScript. A web-based viewer takes the data and displays calculated thermochemistry tables. This demonstration web page may be viewed at http://www.me.berkeley.edu/gri_mech/plus/test4dav.html.

Portal Working Group

The Portal Task Group looked at several competing portal products including open source products such as Apache Jetspeed, Zope, and Insight as well as commercial offerings such as IBM WebSphere, Sun iPlanet, and others. A decision to use Apache Jetspeed based on feature set and cost as well as development-related advantages including:

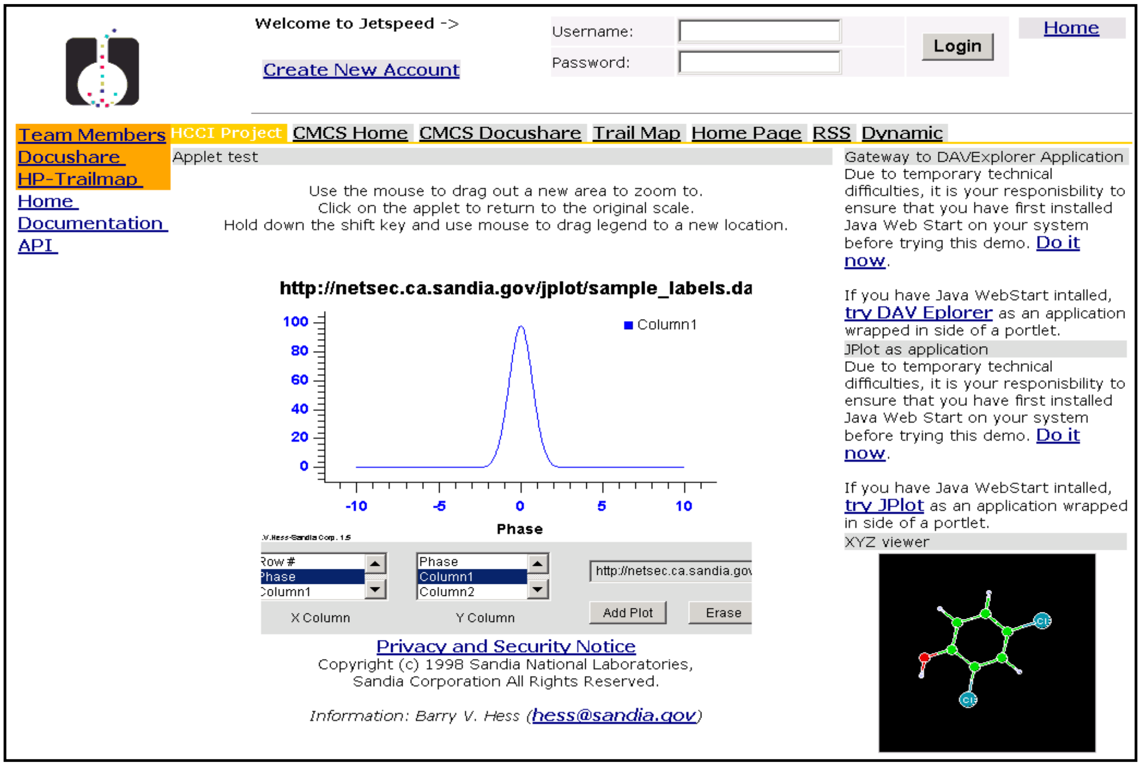
- The ability to examine full source and to influence product direction (Jetspeed is an Apache/Jakarta open source project),
- The potential to leverage efforts across several SciDAC projects and related science driven portal projects (such as the University of Michigan's CHEF effort) that have also chosen Jetspeed,
- Jetspeed's support for the emerging Java portlet standard and it's support for portlet development using Velocity (an Apache/Jakarta templating engine for creating server side pages) and Java Server Pages (JSP)
- The ability to leverage CMCS developers' extensive experience with the Java language and related we development technologies.

A Jetspeed portal prototype (Figure 2) has been installed on the CMCS server that allows use of the tutorial developed by the Portal Task Group. The tutorial demonstrates to the application scientists the current capabilities of Jetspeed including how to access applets, applications, and web pages from within the Jetspeed environment. It also demonstrates how scientists can customize and organize their own pages and panes.

The Portal Task Group has established a number of short term tasks including: setting up a production/test environment for Jetspeed development, evaluating Jetspeed's underlying security model and support for groups/roles, developing a second tutorial aimed at CMCS portal developers; assessing and prototyping collaborative tools as portlets; generating requirements for data access and publishing via the portal; and creating a prototype working portal environment.

The CMCS project's most central requirement involves data/information access and publishing through the portal. Requirements are being generated based on user needs for data access, management, summary, and notification. Individual scientists will be developing applications that are being invoked outside the portal and would interact with DAV server directly; the portal will be a reporting and management window into different aspects of data. The portal will also be playing the role of file manager / document manager window to DAV servers.

Figure 2. Prototype web page in Jetspeed tutorial.



Use the mouse to drag out a new area to zoom to.
Click on the applet to return to the original scale.
Hold down the shift key and use mouse to drag legend to a new location.

Privacy and Security Notice
Copyright (c) 1998 Sandia National Laboratories,
Sandia Corporation All Rights Reserved.

Information: Barry V. Hess (hess@sandia.gov)

Due to temporary technical difficulties, it is your responsibility to ensure that you have first installed Java Web Start on your system before trying this demo. Do it now.

If you have Java WebStart intalled, try DAV Eplorer as an application wrapped in side of a portlet.

Due to temporary technical difficulties, it is your responsibility to ensure that you have first installed Java Web Start on your system before trying this demo. Do it now.

If you have Java WebStart intalled, try JPlot as an application wrapped in side of a portlet.

Infrastructure Working Group

The infrastructure group initially focused on evaluation and selection of a set of tools (Majordomo, MHonArc, DocuShare, Electronic Notebook, website, etc) to help the CMCS operate as a distributed project team. These initial communications tools are helping us in the drafting and discussion of CMCS use cases, process flows, and requirements. In addition, a Software Engineering Plan (SEP) for the CMCS project was created. The CMCS SEP is derived from the Software Systems Engineering Plan that is used at PNNL and has been accepted by the CMCS project team.

This working group has been managing the software configuration environment and has set up a standardized environment enabling software development and deployment for CMCS. In addition, we are supporting CVS for the CFRFS (A Computational Facility for Reacting Flow Science), a BES SciDAC project led by Habib Najm at SNL.

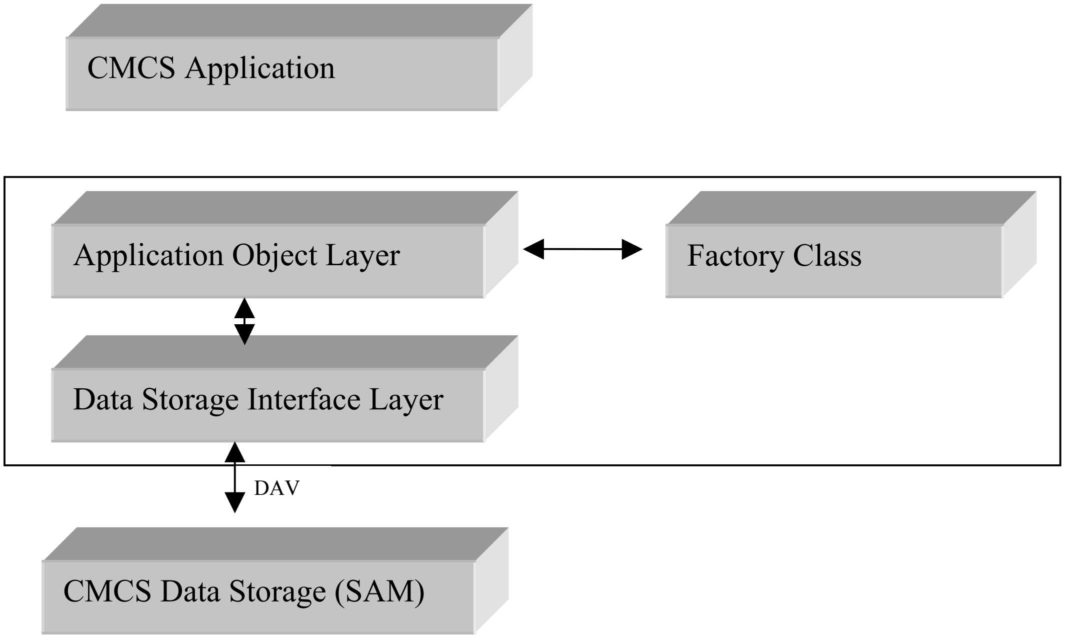
Five task groups have been formed to address the development of CMCS infrastructure: Notification, Searching, Data/Metadata Management, Security, and Pedigree.

The Notification Task Group has developed an initial set of requirements based on use cases and scenario discussions. Notifications on CMCS will typically involve small and infrequent messages that communicate significant events between researchers. Notifications will also be used to communicate between software components and to trigger automatic computations. The extent to which notification will fill this role is still evolving and the requirements will evolve accordingly. Based on the initial requirements, we are looking at several messaging system solutions including an extended Java Messaging System (JMS), JXTA, Grid Events and JAXM. None of these approaches fully meets CMCS requirements. Each has advantages and disadvantages and with the exception of JMS is still evolving. We are monitoring developments with each of these approaches but have chosen to base initial developments on JMS technology.

We have started to test the suitability of OpenJMS from a performance and robustness standpoint. A messaging system with example programs will be made available for experimentation by applications developers by the end of March. The initial API will be Java-based.

An initial set of requirements for "searching" has been developed from both the early proposal process as well as the search process document. Requirements review with other Search Task Group members as well as discussions with application scientists to further understand the desired search targets, search criteria, interaction, and results presentation are underway. Technologies and products that will satisfy the CMCS search requirements and will be compatible with the CMCS data/metadata management infrastructure are being investigated. Technical discussions are held regularly regarding what the data/metadata management infrastructure looks like, what SAM provides, and what will be provided to support searching.

This CMCS Data Management Task Group is working on a wide reaching set of goals to



a
re
(SAM) project. Numerous discussions with Application Scientists have already produced data management concept refinement, artifacts from requirements gathering, and a data storage interface (DSI) applications programming interface (API) document. An API sub-task group that is a composite of Infrastructure software developers and an Application Scientist was spawned recently and work is underway to implement the API in a number of languages to support state of the art code and legacy code. Because the

Distributed Authoring and Versioning (WebDAV) is a relatively new protocol, a set of tutorials were developed and a WebDAV server was setup to familiarize CMCS team members on not only how the protocol works, but also to provide ideas on how CMCS Application Scientists can build and organize their data. This task group has been involved in working in conjunction with the SAM project, the Search Task Group, the Data Pedigree Task Group, the Notifications Task Group, the Portal Task Group, and the XML Task Group to ultimately design the CMCS Data Management Architecture.

The Security Task Group was formed earlier this year to assess CMCS's security requirements. Its initial work has been focused on defining more explicitly who the user base will be, and what resources need to be made secure. Extensive interviews with the project's pilot users (application chemists) have established the primary security requirements, which revolve around the security of the CMCS digital content repository. Given CMCS' intention to use the WebDAV (Web Document Authoring and Versioning) protocol as the technology for storing the CMCS data and metadata content, security will be managed primarily using standard HTTP and DAV mechanisms, e.g. the Secure Sockets Layer (SSL) protocol, username/password authentication, DAV access control lists (ACL). Work is ongoing to understand how migration to public key/grid-based security mechanisms can be accomplished and how the security context can be expanded to include non-DAV resources over the life of the project.

The discussions with the pilot user base have established a robust analogy between publishing papers in the peer-reviewed journals and the publication of data and other content on the CMCS. The security requirements for published CMCS content have been established (content will be digitally signed and time-stamped, and cannot be retracted or altered after publication). Security requirements for works-in-progress, i.e., unpublished data, are being drawn together. These requirements will include group- and role-based permissions for project workspaces that can be administered by project coordinators. This latter effort is being performed in parallel to the development of WebDAV protocols for access controls, in addition to the development of specific WebDAV implementations. In addition, CMCS will be collaborating with the Scientific Annotation Middleware (SAM) project on this problem.

The Pedigree Task Group involves both application scientists and infrastructure team members. It is closely aligned with the Search and Data Management groups as well as the XML Task Group and also the external Center for Collaborative Research in Education (CCRE) Workbench effort. The application scientists have expertise in defining the metadata and determining what pedigree information is necessary. The infrastructure team members have expertise in the implementation of solutions. Data pedigree is at the heart of the CMCS project. It is a relationship that provides a "line of ancestors". This allows for the categorization and tracing of the scientific data, possibly across scales. Scientific pedigree of a piece of data allows identification of its ultimate origin. A pedigree requirements document has been drafted and circulated to the CMCS. A pedigree tutorial had been developed and given to the team. The Pedigree task group, in coordination with the with the Search and Data Management groups, is holding joint requirements gathering discussions with application scientists to determine their data/searching/pedigree needs. A decision has been made to adopt the Dublin Core as an initial set of pedigree metadata within the CMCS. Work is underway to develop a DAV repository that will demonstrate the use of Dublin Core to the entire team.